Current Problems of Using Big Data in Kazakhstan

Serik Aliaskarov, Sholpan Saimassayeva, and Assel Smaiyl

***Abstract*—Currently Kazakhstan suffers from improper use of data analytical techniques and it will play dramatic role for competition among other developing countries. Kazakhstan does not involved into Data mining from large players such as: Google, Facebook and others media and social networks. There are few ICT companies those names will not be mentioned by ethic reasons. One of them has very high potential for developing in that direction, but their niche is only antifraud system for the banks. The second one has some similarity of Big Data Analytics. It has miserable form of making media and social networking analysis for region administrations. However, it shows that there is a demand and potential for BDA. Kazakhstan has a certain level of ICT infrastructure and well working government and public database, those integrated trough e-government GATE. Some barriers make it difficult to interact society and government. Kazakhstan can reach a new level of ICT development by establishment of knowledge of Big Data. It is necessary to emphasize on scientific and practical research to achieve requirements of Bid Data.**

***Index Terms*—Big Data; Google; Facebook; Android; social networking; data-mining; e-government.**

I. Introduction

There is a growing trend of Big Data practice all over the world especially in USA, ICT dominator, where the main players based. Without any doubt, one of the leaders of data mining is Google with their flagship products such as, Android platform for mobile devices, YouTube etc. Another giant player is Facebook. However, let us first take view into Android platform that has around 2 billion users in the world. It means, Google stores an immense amount of information about 26 % of the world's population.

It is very bright example of Bid Data collection because if even exclude simple personal information such as sex, age, email, post address, mobile number, credit card that you must fill during registration. At the same time through different instruments of data mining like Chrome browser, Google search engine, Google Map, Google Play, Google Translator by using analytics layer it is easy to identify full portrait of users. Google tracks and knows where you are actually staying, where you are working, which places you are visiting, traveling, which hotel you are using worldwide, what are you eating, what is your interests, hobbies, they even can identify what is your income, what is you religion, what car are you driving and so on. Moreover, it is just "top of iceberg".

Second key player that mentioned earlier is Facebook that has almost the same number of users – around 2, 07 billion. Social networks have become an integral part of modern society Social networking sites make people share whatever they want and communicate easily. Facebook collects this huge amount of detailed personal information produced by people every single day worldwide. Both of them complement each other.

There are many definitions of Big Data. The scale and challenges of Big Data are usually identified as 4Vs or 5Vs. Gartner has noted four major challenges (the 4Vs): increasing volume of data, increasing velocity, increasing variety of data types and structures, and increasing variability of data. The fifth V is value, which is the contribution big data has to decision-making. (It is shown in Fig. 1.) Add to these the increasing number of disciplines and problem domains where big data is having an impact and one sees an increase in the number of challenges and opportunities for big data to have a major impact on business, science, and government [1], [2].

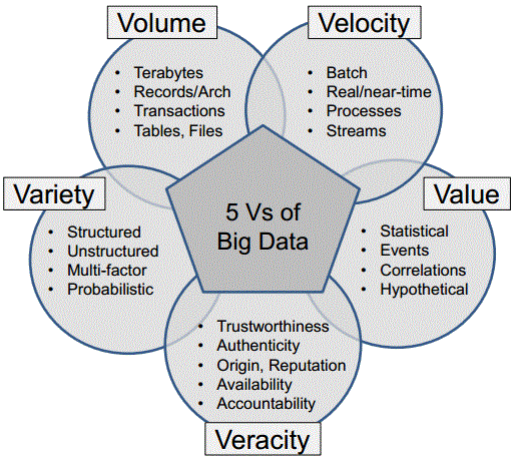

Fig. 1. 5Vs of Big Data.

Kazakhstan in case to compare with other developing countries has very good e-government infrastructure where main player is JCS NIT www.nitec.kz – National Information Technologies, Government Company that provide almost all data bases (structured) such as data base of all citizens, data base of all business (companies, entrepreneurships, and other all types of organizations), data of all addresses and real estate. There are two main principles of above databases are personal identification number and business identification number. Both identification numbers are fully integrated, through e-government GATE with other databases such as; police, court, procurator, tax committee, social and other government databases, it is shown in Fig. 2. Moreover, public and private organizations also integrated through e-government GATE, there are banks, insurance, education, transport etc., so we can identify that Kazakhstan almost ready to use big data technologies. However, there are many

Manuscript received September 27, 2018; revised July 27, 2019.

The authors are with the IITU, Kazakhstan (e-mail: s.aliaskarov@gmail.com, saimassayeva@gmail.com, syrymbayeva.assel@gmail.com).

doi: 10.18178/ijfcc.2019.8.3.545

databases those not ready or non-existing by unknown reasons, like healthcare statistics.

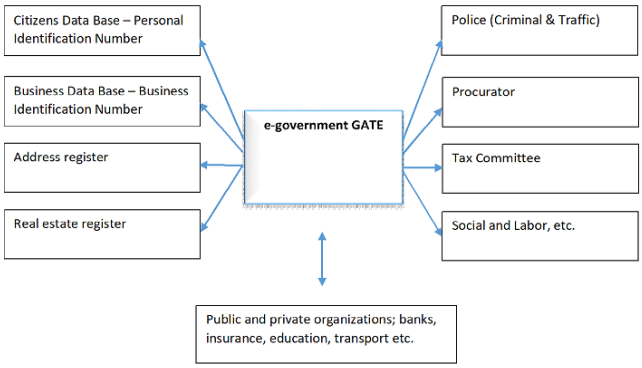

Fig. 2. e-government of Kazakhstan Database Scheme.

Kazakhstan e-gov architecture has 4 levels:

1) Architecture of Activities
2) Architecture of Data
3) Architecture of Systems
4) Architecture of Infrastructure

As a developing country, Kazakhstan faces certain legal obstacle. Government can monitor their citizens with the mosaic effect, while information on government activities remains closed, and many citizens have only limited access to information, edited by their corporate provider.

For a developing country, that does not support certain political co-operation and the continued involvement of public administration, the issue of big data may be very problematic, and it is almost impossible. This is happening in Kazakhstan, which in a certain sense is not so good for us. There is high cooperation between the state and the private sector in many other developing countries. This is something that a country of 18 million people simply cannot do.

II. Problems of Using Big Data in Kazakhstan

As we can find in introduction part, that Kazakhstan has good ICT infrastructure that in some cases can face into few requirements of using Big Data.

The big data paradigm has been designed for achieving the optimal management and analysis of such large quantities of data (big data analytics). The performance of such analytics – possibly influenced by a number of heterogeneous factors such as the type of data, the class of problem to address, or the underlying processing systems – is of the utmost importance as impacting both the effectiveness and the cost of the overall knowledge extraction process [3]. In this context, big data benchmarks are therefore useful to generate application-specific workloads and tests in order to evaluate the analysis processes for data matching [4]. However, the position of Kazakhstan in BDA practices is still poor and there are many things to do to get 5V and make Big Data to serve to society and government.

I mark five main problems such as:

- *Social networks*
- *Search engine and mobile platform*
- *Bid Data platform*
- *Small amount of Databases*
- *Real/near time data*

Social media users lack the ability to configure and enforce customized privacy policies according to the precise privacy measure that they demand. To this end, we argue that there is an emerging need for third-party solutions that are independent of existing social media software and that can detect custom privacy policy violations. In order to address this need, we propose an approach to detecting custom privacy violations for social media users [5].

Social networks – Kazakhstan still does have any specific organization that works with social networks. Currently there are few popular bloggers those use social networks either for own popularity or paid orders to manipulate with public opinion [6]. Problem is that Kazakhstan misses such opportunity to make social network as a platform to make Communication Bridge between society and government and here I mean not only direct communication.

Identification and quantification of influential spreaders in social networks are challenging due to the gigantic network sizes and limited availability of the entire structure. Such difficulty can be overcome by reducing the problem scale to a local one, which is essentially independent of the entire network. This is because in viral spreading the characteristic spreading size does not depend on network structure outside the local environment of the seed spreaders [7].

That can be used to analyze everything that excite and disturb society. It will help to support decision-making process properly in different spheres. 70% of population in Kazakhstan using social networks and logically assume that most of those users average age 16-40 years, it shown in Fig. 3. According to the research, which conducted on16th February till 16th March among the population of Kazakhstan, representing villages and cities at the age of 18 to 55. The survey was conducted by personal interview method, the quantity of respondents N=6054 [8].

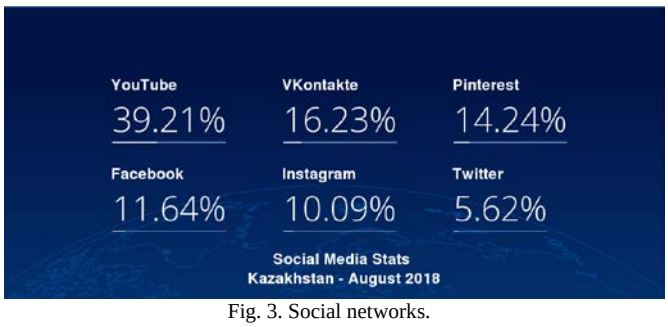

Fig. 3. Social networks.

As it shown in Fig. 3 this graph is important as they illustrate the dimensions of social networks established by social network platforms worldwide [9]. The number of individuals interacting via these platforms is extremely high, and this number is increasing daily [10].

A. Search Engine and Mobile Platform

In Kazakhstan, accordingly to statistics TNS Infratest – connected consumer survey 2016 – 65 % of population have smartphones and 77 % using Internet daily for personal use. Important to emphasize that, 80 % of smartphone users have Android platform, 18 % IOS and Window less than 1 %. However, IOS does not have own search engine, it shown in Fig. 4 [11]. In addition, in Kazakhstan, there is search

engine service from Russian Yandex but it is not popular as in Russia [12].

In Fig. 5 without any doubt, Google is leading in search engine service.

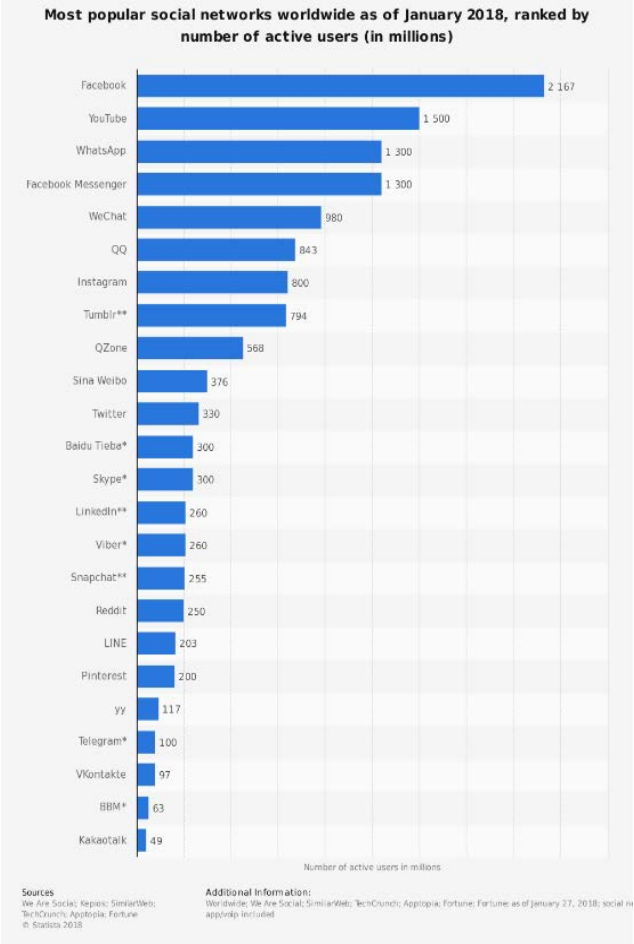

Fig. 4. Social network users.

The same situation with social networking. Probably, there is no company operating with immense amounts of data gathered from searching engine services.

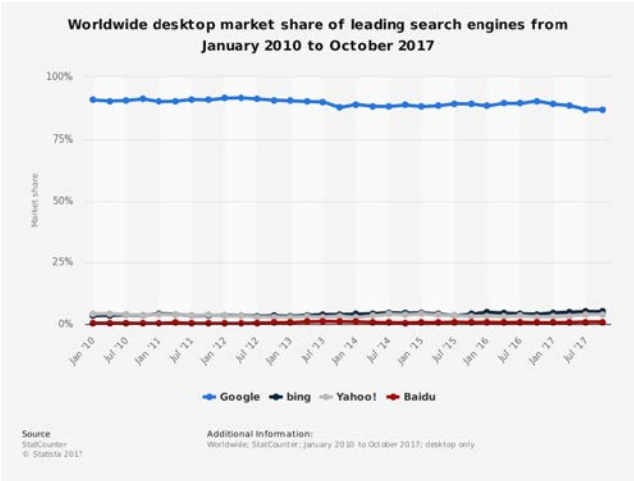

Fig. 5. Search engines.

B. Big Data Platforms

Another problem is that no one in Kazakhstan using Big Data platform and analysis tools. There are three main technological solutions for storage and providing quick access to Big Data:

1) Increasing hardware performance: use faster processors, use more processor cores (requires parallel operations, take advantage of multi-core processors), increase disk space and data transfer capacity, increase network bandwidth (MPP);
2) Reduction of data size: compression of data and data structures, which, by definition, limit the amount of data required for queries (for example, column-oriented databases) (NoSQL) [13];
3) Data distribution and parallel processing: data input on a larger number of disk missives for parallelizing processing and operations, data distribution on separate computing nodes that can work in parallel, using a distributed architecture with a high level of fault tolerance and performance monitoring with a higher network bandwidth for better transmission data between nodes (Hadoop and MapReduce) [14].

MPP (Massively Parallel Processing), NoSQL, Spark, Hadoop and MapReduce are key technologies for working with Big Data. However, Hadoop is main open source player for Bid Data and has many modules such as HDFS, YARN, HBase and so on, from another hand Spark work faster than Hadoop but it does not provide own distributive storage system, it shown in Fig. 6 [15], [16].

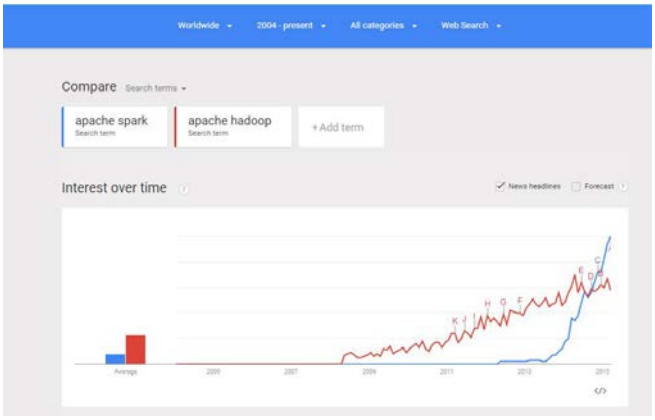

Fig. 6. Image source: Google analytics.

C. Small amount of Databases and real/near time data

As it was mentioned above, Kazakhstan already has reached good results in e-government, has made well-structured data bases, infrastructure and e-government GATE that already providing millions of services to public every day. At the same time, still there is place for bureaucracy among government institutions. One of the good examples is health care. Even if it is one of the most important sectors for public, there is no common database, every medical organization has its own database storage and there is no any integration process. Second, committee of statistics that still makes most of analyses manually. How can Big Data be involved if there is still manual data collection and analysis?

Collection, mining, integration and analysis is live organism that changing every second [17].

Furthermore, there is missing many other databases those play important role in everyday life.

Along with those two main problems, there are many other issues such as weak use of already existing databases

and infrastructure.

III. Conclusion

As we can find that Kazakhstan does not integrated to worldwide competition of Big Data by main five reasons.

Along with those obstacles, Kazakhstan still does not evaluate importance, advantages and benefits of Big Data Analytics. At same time, in other countries government and private sector already actively using BDA for improving quality of life, and building competitive advantages. However, current situation is still not dramatic and has change develop scientific and practical approaches to involve to such challenge. International Information Technology University makes many scientific researches and practical work on few projects in Big Data hopefully university researchers will get satisfied results within two years.

There are few steps that necessary to evaluate Big Data in Kazakhstan:

- Use and integrate to social networks as a tool of communication
- Establish Big Data platform on Hadoop or Spark
- Use and integrate with world leading search engines and platforms
- Increase amount of databases
- Use and integrate to real/near time databases.

References

[1] S. S. Bhanuse, S. D. Kamble, and S. M. Kakde, "Text mining using Metadata for Generation of Side information," in *Proc. 1st International Conference on Information Security & Privacy 2015*, 2016, vol. 78, pp. 807-814.

[2] Iiht Official Blog. (2014). *5Vs of Big Data*. [Online]. Available: http://iihtofficialblog.blogspot.com/2014/07/5-vs-of-hadoop-big-data.htm

[3] R. Han, L. K. John, and J. Zhan, "Benchmarking big data systems: A review," *IEEE Transactions on Services Computing*, vol. 11, no. 3, pp. 580-597, 2018.

[4] V. Persico, A. Pescape, A. Picariello, and G. Sperlí, "Benchmarking big data architectures for social networks data processing using public cloud platforms," *Future Generation Computer Systems*, vol. 89, pp. 98-109, 2018.

[5] M. J. Baeth and M. S. Aktas, "An approach to custom privacy policy violation detection problems using big social provenance data," Concurrency and Computation-Practice & Experience, vol. 30, no. 7, 2018.

[6] U. Can and B. Alatas, "Big social network data and sustainable economic development," *Sustainability*, vol. 9, no. 11, 2017.

[7] ACT. (201AST Kazakhstan Company presents data from a study of the audience of social networks in Kazakhstan. *Social Networks in Kazakhstan.* [Online]. Available: http://kazakhstan.act-global.com/news/%D1%81%D0%BE%D1%86%D0%B8%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B5-%D1%81%D0%B5%D1%82%D0%B8-%D0%B2-%D0%BA%D0%B0%D0%B7%D0%B0%D1%85%D1%81%D1%82%D0%B0%D0%BD%D0%B5/

[8] Y. Hu, S. Ji, and Y. Jin, "Local structure can identify and quantify influential global spreaders in large scale social networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 29, pp. 7468-7472, 2018.

[9] Statista. (2018). Social Media Stats Kazakhstan-August 2018. [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[10] T. D. Chen, S. F. Liu, D. Q. Gong, and H. H. Gao, "Data classification algorithm for data-intensive computing environments," *Eurasip Journal on Wireless Communications and Networking*, no. 219, 2017.

[11] Statista. (2018). Most popular social networks worldwide as of January 2018, ranked by number of active users (in millions). [Online]. Available: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

[12] A. Cuzzocrea, "Semantics meets big data: Formal models, practical issues, novel paradigms," *Journal on Data Semantics*, vol. 5, no. 1, pp. 1-2, 2016.

[13] V. Dhar, M. Jarke, and J. Laartz, "Big data," *Business & Information Systems Engineering*, vol. 6, no. 5, pp. 257-259, 2014.

[14] S. Kaisler, F. Armour, and J. A. Espinosa, "Introduction to big data: Challenges, opportunities and realities," in *Proc. 47th Annual Hawaii International Conference on System Sciences, IEEE*, Waikoloa, HI, 2014, pp. 728-728.

[15] J. Stefanowski, K. Krawiec, and R. Wrembel, "Exploring Complex and Big Data," *International Journal of Applied Mathematics and Computer Science*, vol. 27, pp. 669-679, 2017.

[16] S. Pote. (2016). Spark Vs Hadoop. [Online]. Available: https://muniversity.mobi/blog/spark-vs-hadoop/

[17] H. Tenkanen, E. Di Minin, V. Heikinheimo, A. Hausmann, M. Herbst, L. Kajala, and T. Toivonen, "Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas," *Scientific Reports*, vol. 7, no.17615, 2017.

Aliaskarov Serik was born in Almaty, Kazakhstan on March 3, 1980. His bachelor's degree in engineering and technology and MS's degree in technical sciences were obtained at the JSC KBTU, Almaty, Kazakhstan. He is currently pursuing a doctorate in information systems. Now He is the General Manager of Avaya Kazakhstan and has over 20 years of experience. He created first big data working platform in Kazakhstan. And his research interests are big data in social networks.

Saimassayeva Sholpan was born in Shymkent, Kazakhstan on July 12, 1992. Her bachelor's degree in engineering and technology and MS's degree in technical sciences were obtained at the JSC International IT University, Almaty, Kazakhstan. She is currently pursuing a doctorate in information systems. Now she is a senior lecturer at Department of Information Systems, JSC International Information Technology University. Her research interests are big data and data science.

Smaiyl Assel was born in Almaty, Kazakhstan on January 22, 1992. Her bachelor's degree in engineering and technology and MS's degree in technical sciences were obtained at the JSC International IT University, Almaty, Kazakhstan. She is currently pursuing a doctorate in information systems. Now she is a senior lecturer at Department of Information Systems, JSC International Information Technology University. Her research interests are smart education, big Data, data science.